Nucleic Acids Research, 2004, Vol. 32, Database issue ***D245–D250***
DOI: 10.1093/nar/gkh064

3D-GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes

Keiran Fleming[1], Arne Müller[1,2], Robert M. MacCallum[2] and Michael J. E. Sternberg[1,2,*]

[1]Department of Biological Sciences and Centre for Bioinformatics, Imperial College London, South Kensington Campus, London SW7 2AZ, UK and [2]Biomolecular Modelling Laboratory, Cancer Research UK, 44 Lincoln's Inn Fields, London WC2A 3PX, UK

Received August 8, 2003; Revised and Accepted September 26, 2003

ABSTRACT

The 3D-GENOMICS database (http://www.sbg.bio.ic.ac.uk/3dgenomics/) provides structural annotations for proteins from sequenced genomes. In August 2003 the database included data for 93 proteomes. The annotations stored in the database include homologous sequences from various sequence databases, domains from SCOP and Pfam, patterns from Prosite and other predicted sequence features such as transmembrane regions and coiled coils. In addition to annotations at the sequence level, several precomputed cross-proteome comparative analyses are available based on SCOP domain superfamily composition. Annotations are available to the user via a web interface to the database. Multiple points of entry are available so that a user is able to: (i) directly access annotations for a single protein sequence via keywords or accession codes, (ii) examine a sequence of interest chosen from a summary of annotations for a particular proteome, or (iii) access precomputed frequency-based cross-proteome comparative analyses.

INTRODUCTION

The interpretation and exploitation of the information generated by the genome sequencing projects requires the development of extensive computational tools. Once the set of sequences encoded by the genome (the proteome) is identified, the proteins must be annotated in terms of their structure and biochemical function(s). This involves the use of a range of methodologies including: (i) recognition of sequence motifs/profiles that characterize structure or function, (ii) identification of sequence features such as coiled coils and transmembrane regions, and (iii) the identification of homologous proteins for which a structure and/or function has already been identified or proposed. The recent explosion in the amount of genome sequence data being generated compels the development and maintenance of databases and tools to perform fully automated protein annotation.

There are several databases that provide general annotations of protein sequences, including PEDANT (1), the Genome Information Broker (2) and GeneQuiz (3). There are additionally several specialized databases that provide information regarding the possible protein folds and structures for protein sequences, e.g. SUPERFAMILY (4), GENE-3D (5), SMART (6) and GTOP (7). A major difference between 3D-GENOMICS and these databases is the facility for generating dynamically constructed tables of statistics of SCOP superfamily composition between proteomes. Static tables are available from other resources but they cannot be customized like those available in 3D-GENOMICS.

Often there is a degree of duplication in the information contained within the existing databases. However, these databases provide not only different views of the same data but also novel information due to different annotation procedures that highlight consensus and disagreement [for example, we have benchmarked and optimized the use of PSI-BLAST for the recognition of remote homologues with <20% sequence identity (8)]. The existence of several databases that have attempted similar annotation approaches should be regarded as a benefit for this reason, with different implementations, or alternative combinations, of software leading to a different focus of the database. Indeed, the Interpro resource (9) recognizes that databases have different areas of optimum application and so combines several of the aforementioned databases to provide a consistent view of protein signatures.

Here, we introduce 3D-GENOMICS (http://www.sbg.bio.ic.ac.uk/3dgenomics/), a database that provides a broad range of structural and functional annotations for protein sequences from sequenced genomes. The database is proteome oriented

*To whom correspondence should be addressed at: Biochemistry Building, Department of Biological Sciences, Imperial College London, South Kensington Campus, London SW7 2AY, UK. Tel: +44 20 7594 5212; Fax: +44 20 7594 5264; Email: m.sternberg@ic.ac.uk
Present addresses:
Arne Müller, Aventis Pharma, Drug Safety Evaluation, 13 quai Jules Guesde, 94403 Vitry-sur-Seine Cedex, France
Robert M. MacCallum, Stockholm Bioinformatics Centre, Department of Biochemistry and Biophysics, Stockholm University, S-106 91 Stockholm, Sweden

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

and is intended both as a portal to the state of annotations on individual protein sequences, and as a platform for structure-based comparative analyses of proteomes, combining both types of database discussed above. 3D-GENOMICS has a particular focus on utilizing structural annotation. This is achieved through the inclusion of the PDB (10), SCOP (via the ASTRAL database) (11,12) and Pfam (13) databases from which domain information and common ancestry can be inferred. This focus has been chosen because protein domain-based annotations generally provide more detail than those based on the protein sequence as a unit (14). Furthermore, the use of structural domains may prove a valuable tool for the biologist to design bench experiments (e.g. locating surface-exposed residues on the protein for site-directed mutagenesis experiments). Consideration of common protein structure between sequences additionally allows detection of more distant homologous relationships between proteins than would be available from sequence-based analysis alone, due to the fact that 3D structure is often more conserved than sequence (11,15).

ANNOTATION PIPELINE

We have implemented an automated pipeline to generate the annotations, which are stored in a MySQL database. The pipeline is split into several steps as shown in Figure 1 (the numbers in the figure correspond to the steps listed below), starting with a single sequence and finishing with a comparative analysis of proteomes.

1. Basic sequence features are assigned to protein sequences. These include the identification of homologous sequences via BLAST and PSI-BLAST (16), IMPALA (17) and HMMer (18) from different databases such as SwissProt (19), PIR (20), Pfam, PDB and SCOP. Conserved sequence patterns from the Prosite (21) database are assigned to sequences. The database also contains predictions of transmembrane helices via HMMTOP (22), coiled coils via COILS2 (23), regions of biased amino acid composition via SEG (24) and secondary structure via PSI-PRED (25). The numbers of basic sequence features generated are huge; e.g. the human proteome currently contains ~30 000 identified protein sequences, but these account for more than 10 million PSI-BLAST alignments stored in the database.

2. Alignments are merged according to the regions they span within the query sequence and the nature of the homologues. The different types are: (i) domains of known structure from SCOP, (ii) full-chains from PDB entries (i.e. sequences of known structure), (iii) homologues of potentially known function from SwissProt, PIR or PDB that do not contain any of the words 'hypothetical', 'probable', 'putative' or 'potential' in their description, and (iv) any homologous sequences, including conserved hypotheticals and sequences of unknown function. This can be seen in Figure 2, where regions are displayed directly beneath the query sequence. The PDB hit, for example, represents a highly significant PSI-BLAST hit to a PDB sequence spanning as much of the query sequence as possible. In this way, users can quickly visualize important matches without having to search through all of the individual alignments. The original alignments are retained for visualization by the user, yet the purpose of this step is to reduce the amount of data to a few regions per query sequence, and thus to speed up the downstream cross-proteome analyses. For example, this step reduces the number of alignments to less than 81 000 overlapping regions for the human proteome without reducing the information content markedly.

3. The data from steps 1 and 2 are processed to generate a genome-wide statistical summary of the different annotation features. Figure 3 shows an example annotation summary for the Human genome [from ENSEMBL version 7.29a2 (26)] and represents a 'top-down' approach to accessing the database. The numbers of sequences, residues and regions (merged alignments) that can be assigned to each type of annotation are summarized. Also computed are statistics including summaries of the number of proteins containing SCOP classes, folds and superfamilies, and Pfam domains.

4. Several types of comparative analysis are computed in order to provide cross-proteome statistics. Commensurate with our structurally biased research interests, we provide various SCOP-based comparisons between the annotations in the proteomes. Those currently available represent a subset of possible comparisons. Detailed comparative analyses have been published by the authors (27).

USER WEB INTERFACE

There are four major entry points to the database, reflecting the different kinds of question that a user may want to answer:

(i) Keyword search: this option provides a facility for the user to search one or more proteomes for sequences that have an annotation containing the text of interest, e.g. one may want to search for all proteins in human or *Drosophila* that are associated with 'recombination and repair'. The search can be configured in a number of ways, e.g. fuzzy match versus exact match, to allow the user to focus on sequences of particular interest.

(ii) Accession code search: if the accession number from one of the common databases for a given sequence is known by the user, this option allows direct access to that sequence. Alternatively, the user can paste in the one-letter amino acid sequence for the protein and the database will return the sequence if a match is found.

(iii) Top-down analysis: this option provides a list of the annotated proteomes and the user may choose one of several types of precomputed data. Included in this section is an overview of the status of annotation (see Fig. 3), as well as tables of SCOP classes, folds and superfamilies, and a list of Pfam families found in the proteome (see above). Access to the upstream annotations is provided through hyperlinks in the results tables.

(iv) Comparative proteomics: various examples of comparisons based on the frequencies of SCOP domain superfamilies between organisms are available (see below).

Enquiries via (i) and (ii) retrieve one or more sequences from the database. Figure 2 shows an example of the kind of report that is generated for a protein sequence. On this page, all information available in the database regarding this sequence is presented. Included are the accession codes and descriptions from the primary source databases, followed by a graphical alignment of the primary amino acid sequence and all of the annotation features. Homologous sequences from specific databases/proteomes/taxonomic groups can be searched for, with a great degree of customization available

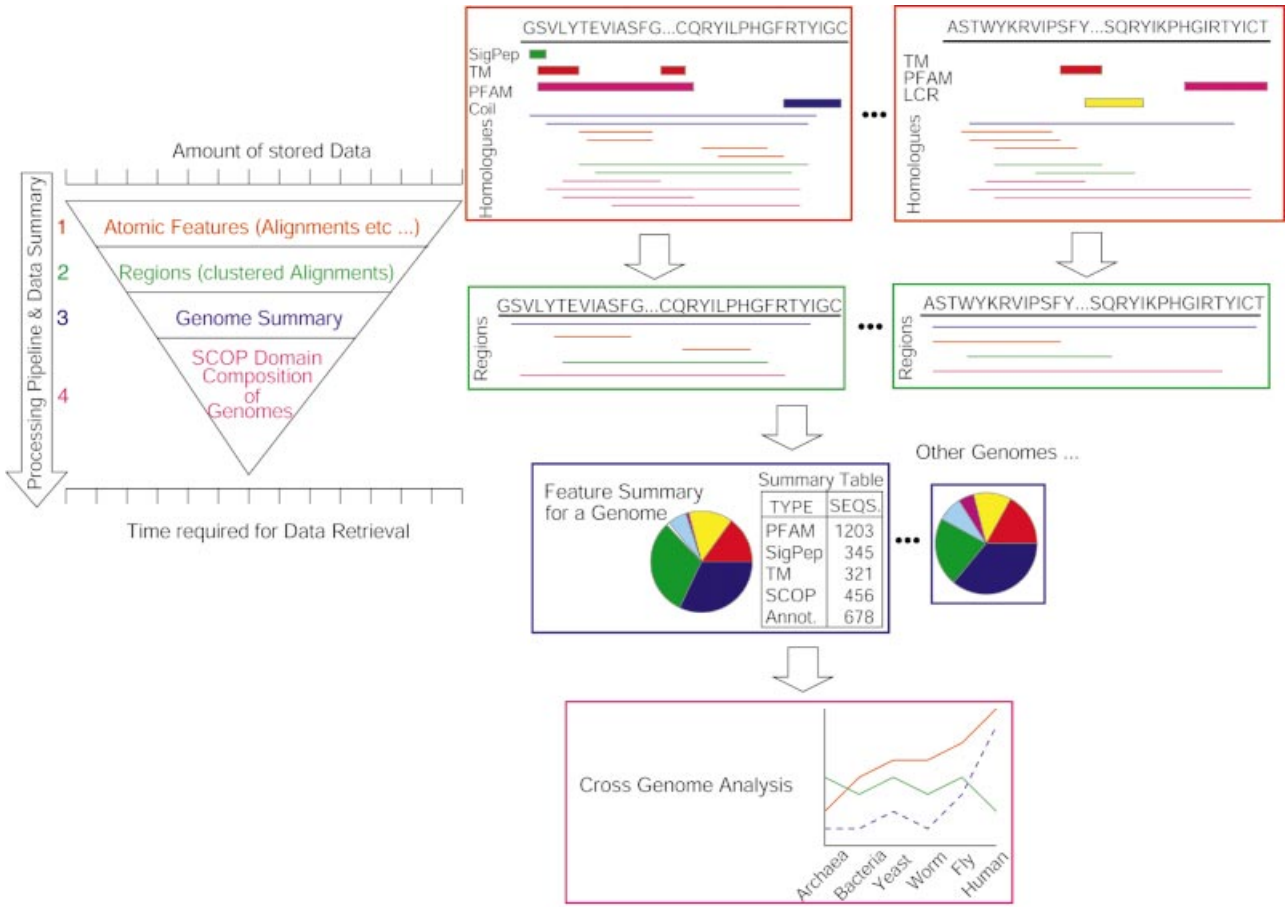


Figure 1. Overview of the annotation pipeline. The triangle on the left symbolizes the reduction in the amount of data at each step in the annotation pipeline. Reducing the data ensures that user database queries return results quickly, and that the results can be viewed at different levels of complexity. The right-hand part of the figure shows examples of the types of annotations produced within each step. The user may query the database, via the web interface, at any of the numbered steps. It is possible to enter the system at the tip of the triangle (step 4) and to follow the annotations (via hyperlinks) back to the basic sequence features, with an increasingly detailed amount of information presented at each stage.

for these searches. For example, homologous sequences can be extracted at a given sequence identity or e-value (a measure of statistical significance of the alignment) threshold or that match a given region within the query sequence. Hyperlinks to the sequences and source databases allow the user to further investigate specific annotations.

Enquiries via (iii) and (iv) provide a 'top-down' approach, where the results generated take the form of proteome-wide or cross-proteome summary tables (a typical proteome-wide summary is shown in Fig. 3). From such an overview the user can retrieve the corresponding list of sequences or domains, i.e. access the 'lower' levels of annotations.

COMPARATIVE ANALYSES

Three types of cross-proteome analysis are currently available. The first provides a comparison of the annotation status for all of the genomes in the database in one table. The second, 'Comparison of SCOP domains ...', enables the user to generate dynamically constructed tables of comparative statistics of SCOP superfamily domain composition between multiple proteomes using the precompiled information described in step 4 of the annotation pipeline. The user has a number of domain properties to choose from, including domain frequencies, normalized SCOP superfamily frequencies (i.e. the number of domains within a superfamily over all domains found in the proteome), the number of sequences that contain at least one domain of a certain superfamily, the average repetitiveness of a domain within a proteome, the number of domains within a superfamily that are found in the globular part of membrane proteins and the repertoire of SCOP superfamilies and Pfam families that are found together with a certain 'partner' SCOP superfamily in the same sequence (i.e. domain co-occurrence) (28–30). The results tables can also be plotted graphically.

The third type of comparative analysis, 'tabulated SCOP venn tables', allows the user to generate reports detailing the union and intersection of SCOP superfamilies found within proteomes. That is, those SCOP superfamilies that are found in all of the selected proteomes, a subset of proteomes, or that are only found in one of the proteomes. This type of analysis may be particularly useful for evolutionary studies (11).

FUTURE DEVELOPMENTS

Maintenance and development of 3D-GENOMICS is now supported via a Biotechnology and Biological Sciences Research Council/Department of Trade and Industry

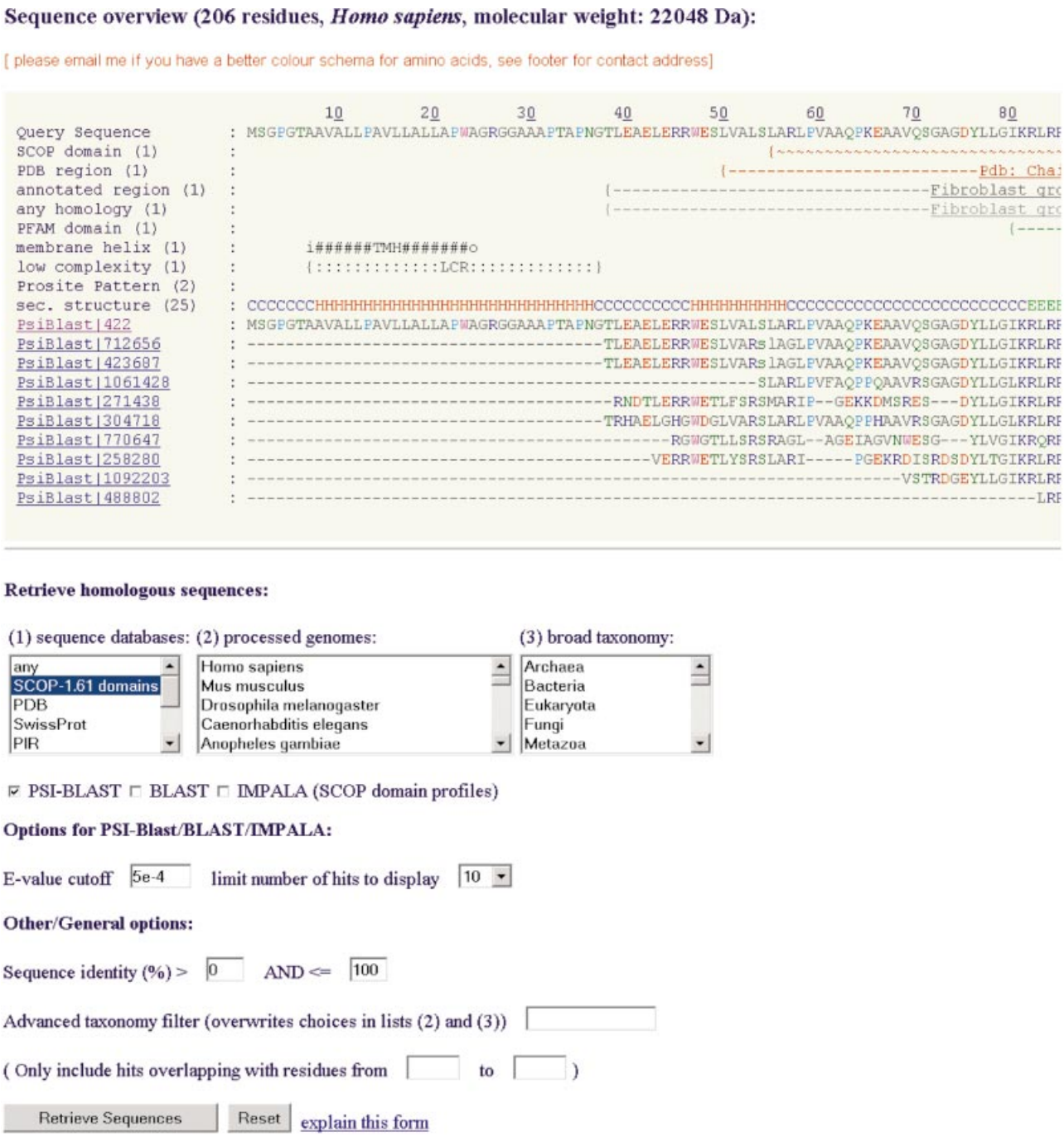
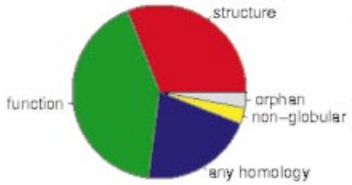


Figure 2. Sequence annotation report for a processed protein sequence. Homologues identified by PSI-BLAST can be requested via the lower part of the window ('Retrieve homologous sequences'). Short descriptions of the homologues are displayed when pointing the mouse cursor over the sequence link (blue). Annotation summaries (regions) are displayed directly below the sequence, for example a SCOP domain starting at residue 55 within the query. Annotations are hyperlinked to their source databases. The sequence report offers more information that cannot be displayed because of limited space.

(BBSRC/DTI) research grant as part of the e-Protein project (http://www.e-protein.org). Recent integration of 3D-GENOMICS into the e-Protein project has ensured a much higher throughput of the annotation procedure via the use of a greatly increased set of resources at three sites: Imperial College London, University College London and the European Bioinformatics Institute. As such, although the annotations currently available are for sequences downloaded in July 2002 (predating involvement in the e-Protein project), recent prototyping of the annotation pipeline using Grid technology (31) suggests that the 3D-GENOMICS database can be updated every 3–6 months. One of the key problems maintaining a database that performs cross-proteome comparative analyses is that incremental updates are almost

Summary of Annotation:

Cumulative residue counts for the basic sequence features:



Feature	Residues	Fraction (%)
Structure	3708411	31
known Function	5052138	42
any Homology	2496810	21
Non-Globular	321626	3
Orphans	382859	3

Annotation Features:

	non-cumulative		cumulative		non-cumulative		cumulative		
Feature	**Sequences**	**Seq.(%)**	**Sequences**	**Seq.(%)**	**Residues**	**Res.(%)**	**Residues**	**Res.(%)**	**Regions**
SCOP	11072	42	11072	42	2541561	21	2541561	21	17635
PDB	13108	50	2330	9	3582675	30	1166850	10	16174
PFAM (known function)	16626	63	4394	17	4130795	35	1646321	14	38966
functional Annotated	20088	76	2744	10	8622831	72	3405817	28	21491
PFAM (unknown function)	118	0	41	0	25307	0	10650	0	129
any Homology	24648	93	4097	16	11210279	94	2486160	21	25468
Transmembrane helix	5484	21	181	1	433459	4	17159	0	19377
Coiled-Coil	3384	13	36	0	229503	2	8713	0	7772
Low Complexity Region	19277	73	900	3	1148592	10	295754	2	60345
[Repeats]	4077	15	-	-	959923	8	-	-	7073
Sum	-	-	25795	98	-	-	11578985	96	-

('?': sequence was not tested for this feature, '-': cannot be calculated, categories in square braces '[]' are ignored for the caclulation of Sum)

Figure 3. Summary of the annotation status of the human proteome (from ENSEMBL version 7.29a2). The pie chart gives an overview of how much of the proteome (in residues) can be assigned to different types of homologue, e.g. SCOP or Pfam domains. Orphans are defined as non-conserved regions without any homology to regions in other proteomes. The table gives details about the number of sequences, residues and regions (merged alignments) that can be assigned to the different annotation categories. Specific sequences and their annotations can be accessed via the hyperlinks in the 'non-cumulative sequences' column, which returns a list of sequences in these annotation categories. The cumulative column provides a running total of the numbers that can be assigned to the different types of annotation, from structural (i.e. SCOP and PDB), through functional [i.e. Pfam (known function) and functionally annotated], to non-globular regions (i.e. transmembrane helices, coiled coils, low complexity regions and repeats).

impossible due to the need for re-computation of these analyses each time new sequences enter the database. Therefore, it is imperative that we significantly speed up our annotations through the use of Grid technology.

Expansion of the types of cross-proteome analysis available is a high priority and will be dealt with shortly. New tools, e.g. to predict disordered regions in proteins (32), will be integrated in the near future, as will new data sources to provide more complete coverage of protein sequences.

REFERENCES

1. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
2. Fumoto,M., Miyazaki,S. and Sugawara,H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res.*, **30**, 66–68.
3. Hoersch,S., Leroy,C., Brown,N.P., Andrade,M.A. and Sander,C. (2000) The GeneQuiz web server: protein functional analysis through the web. *Trends Biochem. Sci.*, **25**, 33–35.
4. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
5. Buchan,D.W., Rison,S.C., Bray,J.E., Lee,D., Pearl,F., Thornton,J.M. and Orengo,C.A. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. *Nucleic Acids Res.*, **31**, 469–473.
6. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
7. Kawabata,T., Fukuchi,S., Homma,K., Ota,M., Araki,J., Ito,T., Ichiyoshi,N. and Nishikawa,K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.*, **30**, 294–298.
8. Muller,A., MacCallum,R.M. and Sternberg,M.J. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.
9. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
10. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.

11. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
12. Chandonia,J.M., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res.*, **30**, 260–263.
13. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
14. Copley,R.R., Doerks,T., Letunic,I. and Bork,P. (2002) Protein domain analysis in the era of complete genomes. *FEBS Lett.*, **513**, 129–134.
15. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
18. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
19. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
20. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
21. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
22. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
23. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
24. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
25. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
26. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
27. Muller,A., MacCallum,R.M. and Sternberg,M.J. (2002) Structural characterization of the human proteome. *Genome Res.*, **12**, 1625–1641.
28. Apic,G., Gough,J. and Teichmann,S.A. (2001) An insight into domain combinations. *Bioinformatics*, **17**, *Suppl. 1*, S83–S89.
29. Apic,G., Gough,J. and Teichmann,S.A. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.*, **310**, 311–325.
30. Bashton,M. and Chothia,C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.*, **315**, 927–939.
31. Foster,I. (2003) The Grid: computing without bounds. *Sci. Am.*, **288**, 78–85.
32. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.